Bioinformatics

CS300

Blast, Substitution Matrices and Protein Alignments

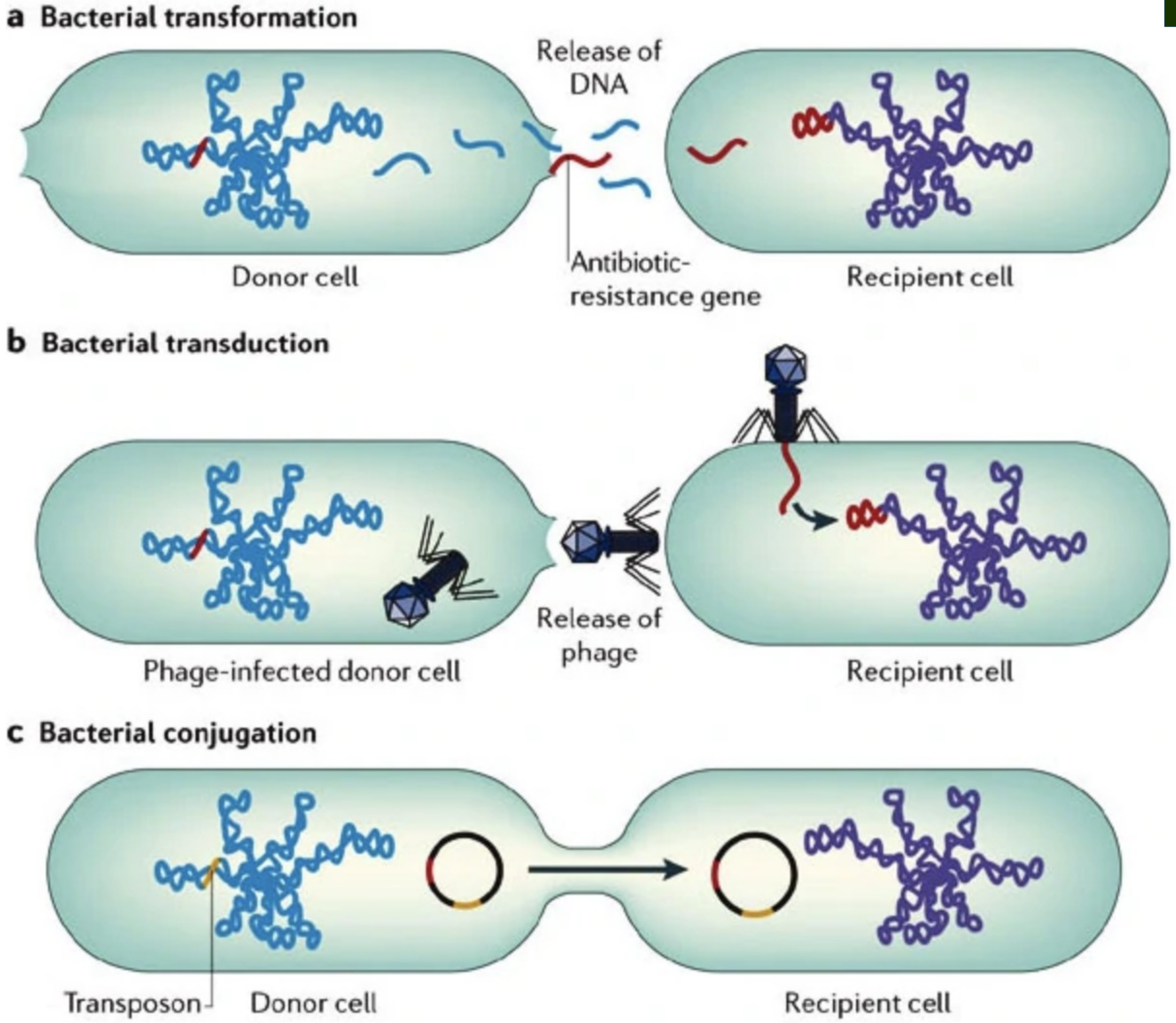
(Chap 4 and 5 in textbook)

Week8, Deck 1
Fall 2022
Oliver BONHAM-CARTER

Horizontal Gene Transfer

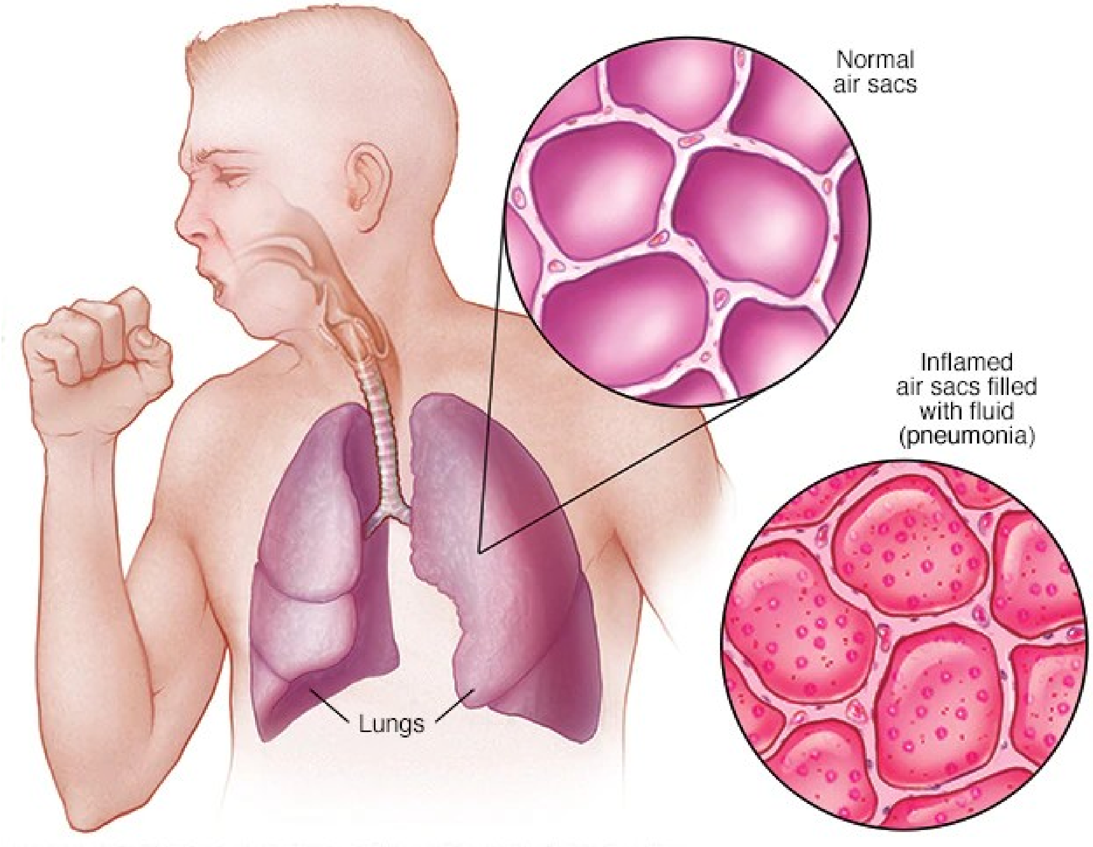
- **Horizontal gene transfer (HGT)**:
  - The movement (transfer) of genetic information between (unrelated) organisms
  - Not generally *relational* genetics (i.e., parent to offspring)
  - Superbugs: A process that includes the spread of genes o for antibiotic resistance among bacteria (except for those from parent to offspring)
  - Fueling pathogen evolution

Methods of Gene Transfer



Pneumonia, *for example*

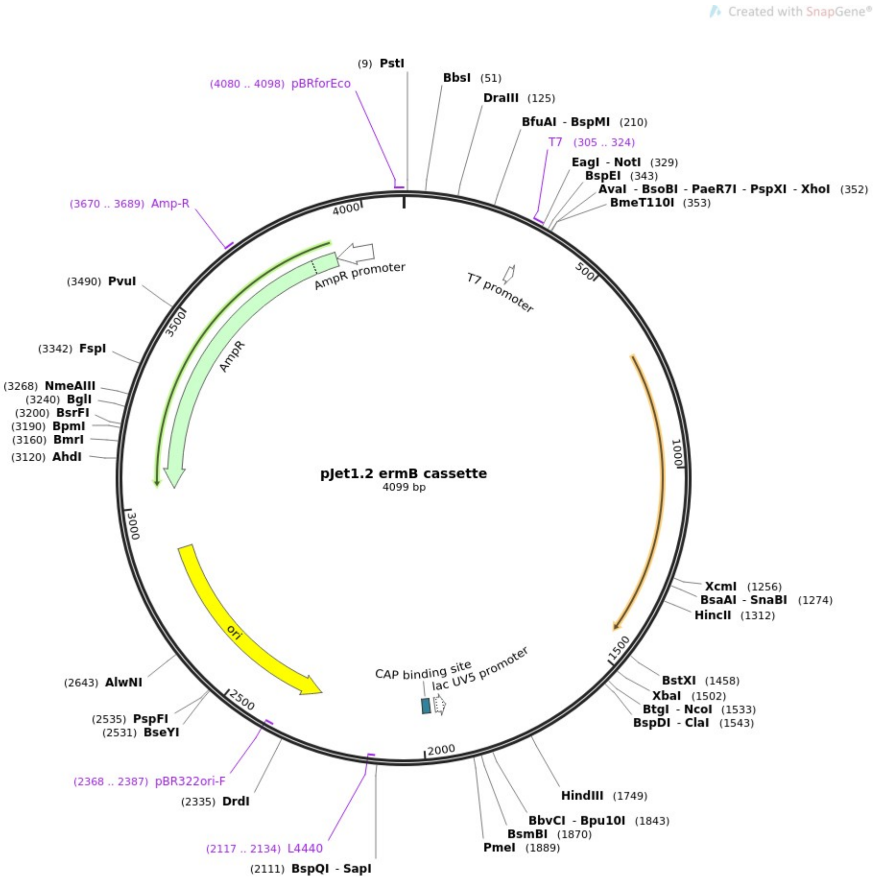
- Pneumonia is an infection that inflames the air sacs in one or both lungs. The air sacs may fill with fluid or pus (purulent material), causing cough with phlegm or pus, fever, chills, and difficulty breathing. A variety of organisms, including bacteria, viruses and fungi, can cause pneumonia.
- A classic sign of bacterial pneumonia is a cough that produces thick, blood-tinged or yellowish-greenish sputum with pus.



https://www.mayoclinic.org/diseases-conditions/pneumonia/symptoms-causes/syc-20354204

Human Pathogen Inquiry:
The *ermB* gene

- An erythromycin-resistance gene from Streptococcus agalactiae, a gram-positive bacterial species commonly associated with the udders of cows, causing mastitis (i.e., inflammation of breast tissue that sometimes involves an infection and may cause fever)



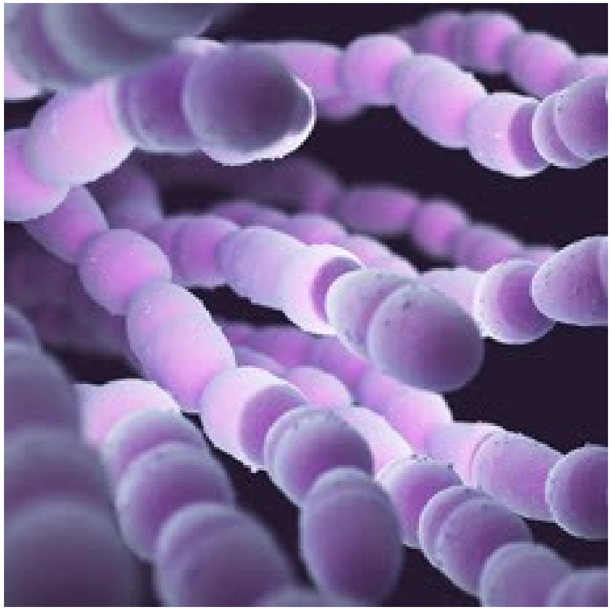
https://www.addgene.org/117121/

Pneumonia and *ermB*

- **Drug resistant**: Erythromycin is a *macrolide antibiotic* (i.e., a drug used to treat various bacterial infections)
- Resistance is due to the *ermB* gene which has been noted in the bacteria, *Streptococcus pneumonia* – a common cause of bacterial **pneumonia**.

Pneumococci

- A type of streptococcus bacteria
- The bacteria spread through contact with illness or by contact with healthy people who carry the bacteria in the back of the nose.
- Pneumococcal infections can be mild or severe.



Horizontal Gene Transfer?

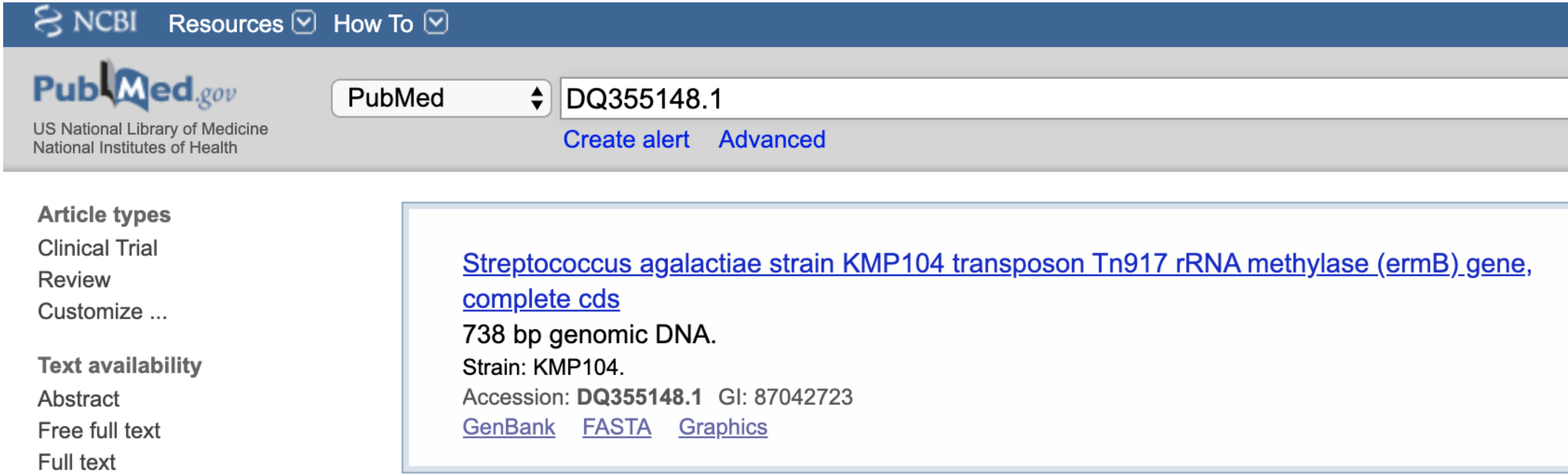
- This type of pneumonia is not believed to have always been resistant to drugs.
- Could the resistance gene have come from another bacteria type via HGT?
- Given all the sequences to analyze, *how could we check what other bacterial organisms have a specific allele for the gene that effectively resists drugs for pneumonia*?
- We will use Blast for this task.

BLAST BLAST BLAST

Let’s Study HGT

- Locate the ***Nucleotide*** sequence for *Streptococcus agalactiae (DQ355148.1) using* https://www.ncbi.nlm.nih.gov/

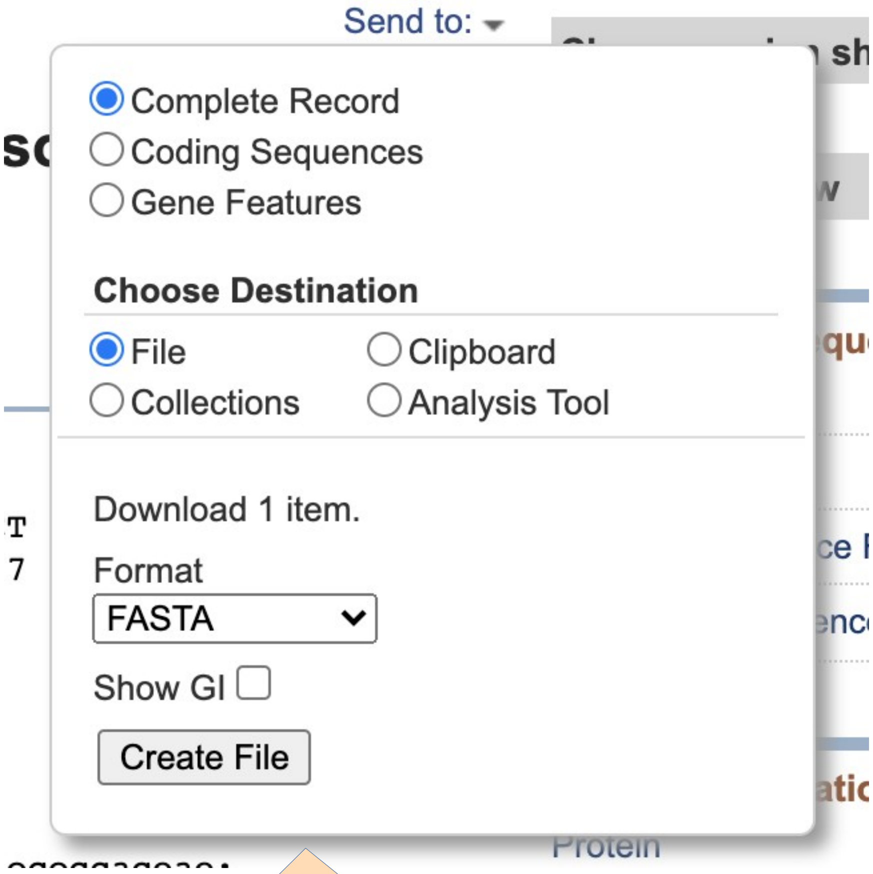
Quick link:
https://www.ncbi.nlm.nih.gov/search/all/?term=DQ355148.1



DQ355148.1

Streptococcus agalactiae strain KMP104 transposon Tn917 rRNA methylase (ermB) gene, complete cds
738 bp genomic DNA.
Strain: KMP104.
Accession: **DQ355148.1** GI: 87042723
GenBank FASTA Graphics

How to get the Data?



Send to: ▾

● Complete Record
○ Coding Sequences
○ Gene Features

Choose Destination

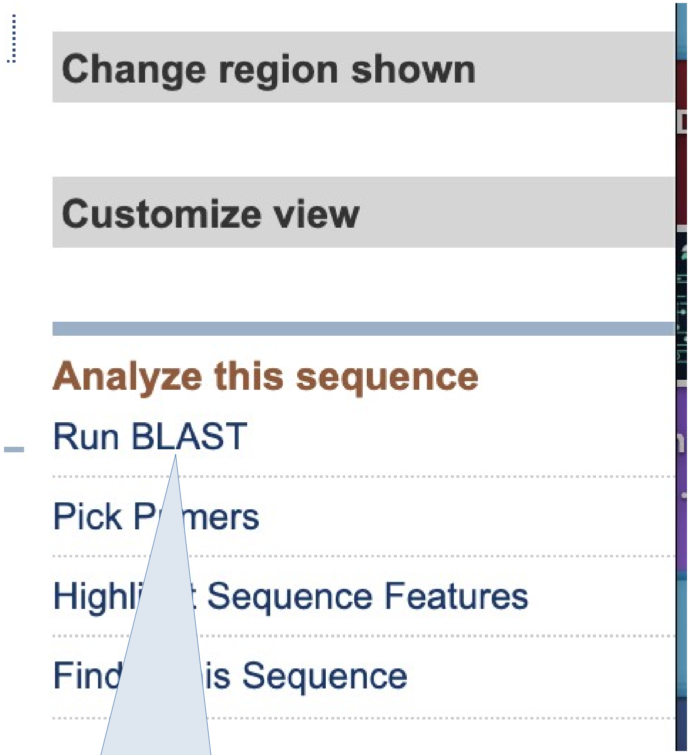
● File ○ Clipboard
○ Collections ○ Analysis Tool

Download 1 item.

Format: FASTA

Show GI □

Create File



Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick P mers

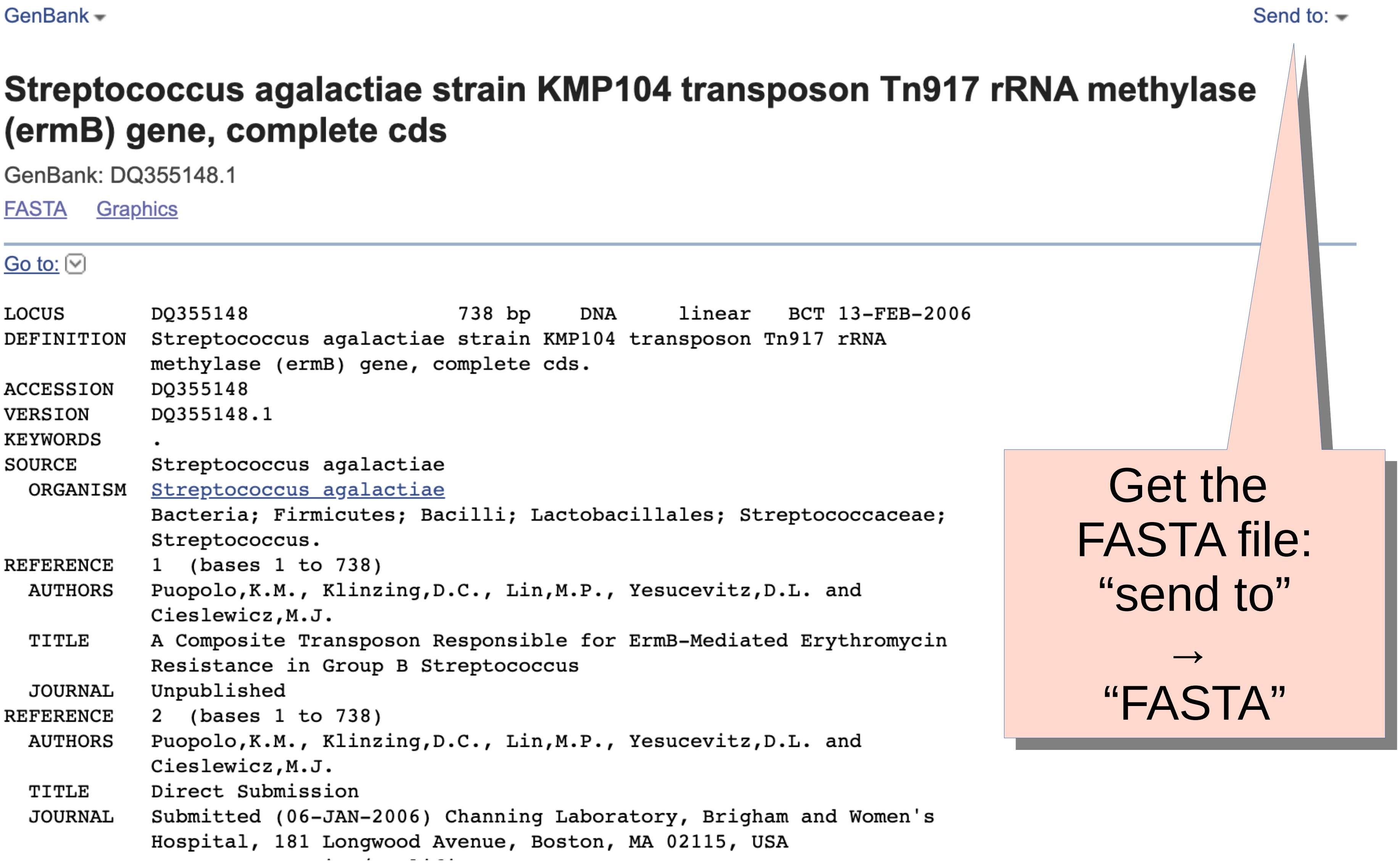
Highli t Sequence Features

Find is Sequence

Method 1:
Get a text file of the gene to have the sequence or now and future work.

Method 2:
Locate a gene record on NCBI and click the Blast button.

Find the Nucleotide Sequence



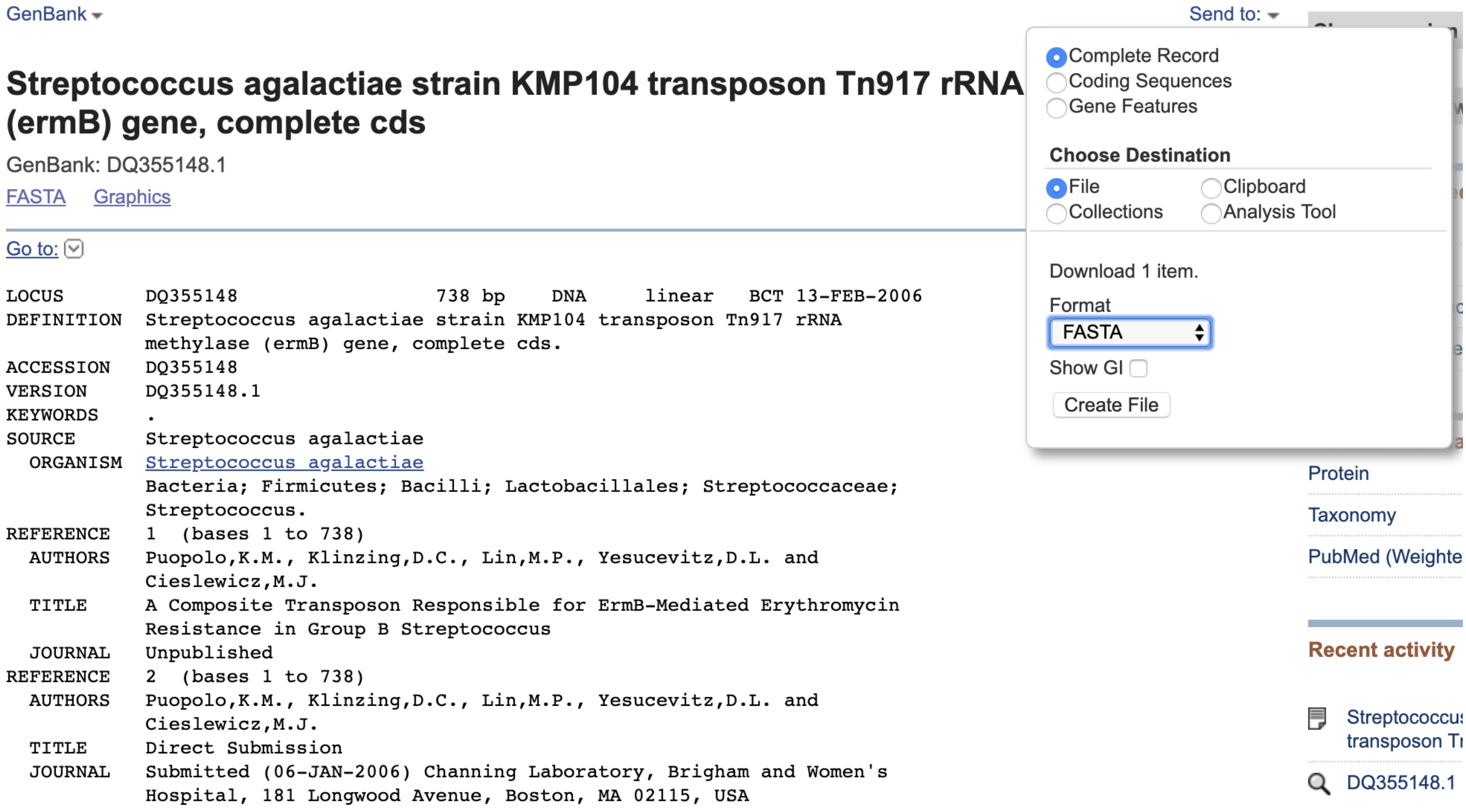
Streptococcus agalactiae strain KMP104 transposon Tn917 rRNA methylase (ermB) gene, complete cds

GenBank: DQ355148.1

```
LOCUS       DQ355148                 738 bp    DNA     linear   BCT 13-FEB-2006
DEFINITION  Streptococcus agalactiae strain KMP104 transposon Tn917 rRNA
            methylase (ermB) gene, complete cds.
ACCESSION   DQ355148
VERSION     DQ355148.1
KEYWORDS    .
SOURCE      Streptococcus agalactiae
  ORGANISM  Streptococcus agalactiae
            Bacteria; Firmicutes; Bacilli; Lactobacillales; Streptococcaceae;
            Streptococcus.
REFERENCE   1  (bases 1 to 738)
  AUTHORS   Puopolo,K.M., Klinzing,D.C., Lin,M.P., Yesucevitz,D.L. and
            Cieslewicz,M.J.
  TITLE     A Composite Transposon Responsible for ErmB-Mediated Erythromycin
            Resistance in Group B Streptococcus
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 738)
  AUTHORS   Puopolo,K.M., Klinzing,D.C., Lin,M.P., Yesucevitz,D.L. and
            Cieslewicz,M.J.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-JAN-2006) Channing Laboratory, Brigham and Women's
            Hospital, 181 Longwood Avenue, Boston, MA 02115, USA
```

Get the FASTA file: “send to” → “FASTA”

Save the Sequence



● Complete Record
○ Coding Sequences
○ Gene Features

Choose Destination

● File ○ Clipboard
○ Collections ○ Analysis Tool

Download 1 item.

Format: FASTA

Show GI □

Create File

Streptococcus agalactiae strain KMP104 transposon Tn917 rRNA (ermB) gene, complete cds

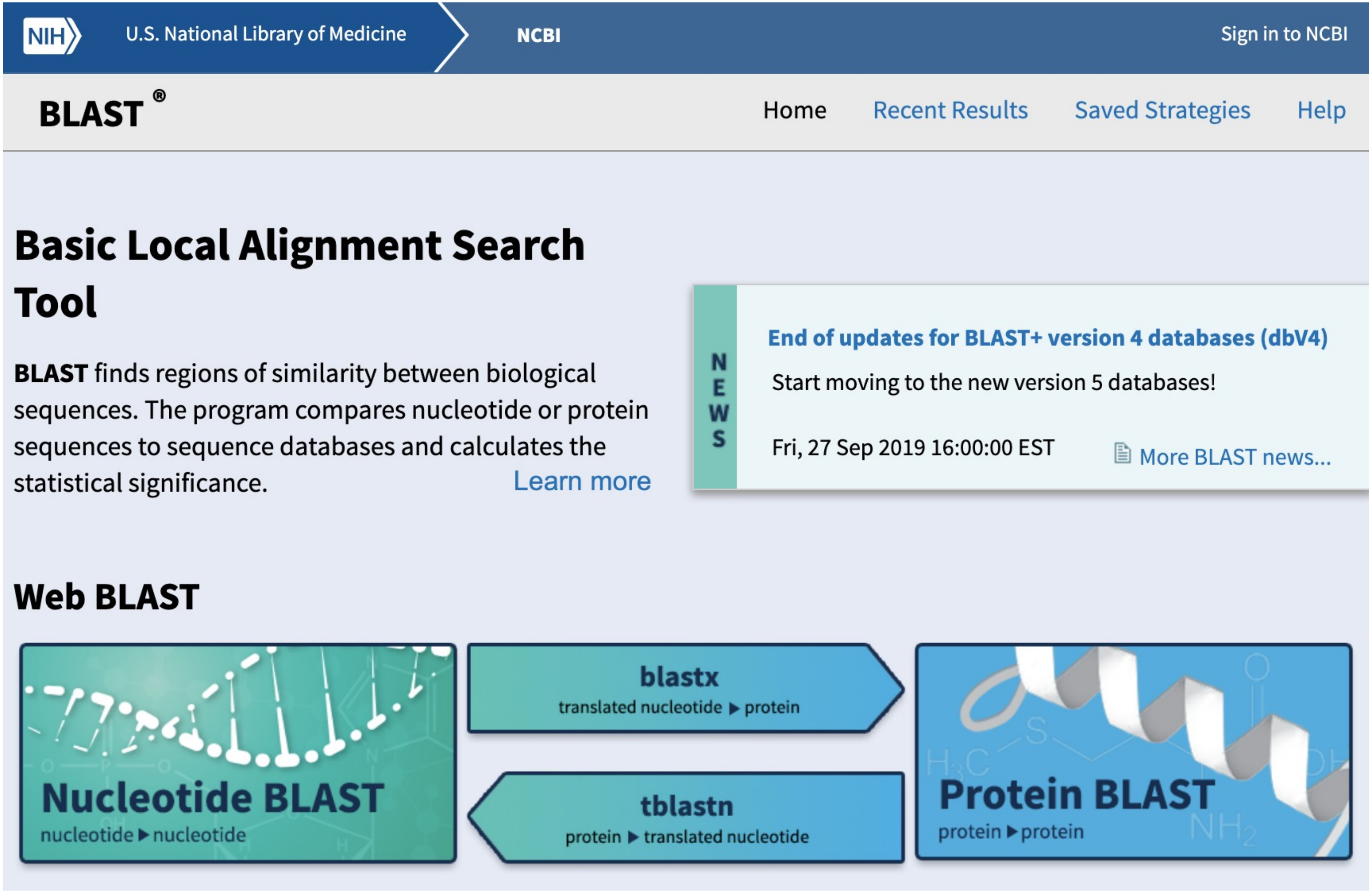
GenBank: DQ355148.1

```
LOCUS       DQ355148                 738 bp    DNA     linear   BCT 13-FEB-2006
DEFINITION  Streptococcus agalactiae strain KMP104 transposon Tn917 rRNA
            methylase (ermB) gene, complete cds.
ACCESSION   DQ355148
VERSION     DQ355148.1
KEYWORDS    .
SOURCE      Streptococcus agalactiae
  ORGANISM  Streptococcus agalactiae
            Bacteria; Firmicutes; Bacilli; Lactobacillales; Streptococcaceae;
            Streptococcus.
REFERENCE   1  (bases 1 to 738)
  AUTHORS   Puopolo,K.M., Klinzing,D.C., Lin,M.P., Yesucevitz,D.L. and
            Cieslewicz,M.J.
  TITLE     A Composite Transposon Responsible for ErmB-Mediated Erythromycin
            Resistance in Group B Streptococcus
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 738)
  AUTHORS   Puopolo,K.M., Klinzing,D.C., Lin,M.P., Yesucevitz,D.L. and
            Cieslewicz,M.J.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-JAN-2006) Channing Laboratory, Brigham and Women's
            Hospital, 181 Longwood Avenue, Boston, MA 02115, USA
```

Ah, The Sequence in FASTA Format

```
>DQ355148.1 Streptococcus agalactiae strain KMP104 transposon Tn917 rRNA methylase (ermB) gene, complete cds
ATGAACAAAAATATAAAATATTCTCAAAACTTTTTAACGAGTGAAAAAGTACTCAACCAAATAATAAAAC
AATTGAATTTAAAAGAAACCGATACCGTTTACGAAATTGGAACAGGTAAAGGGCATTTAACGACGAAACT
GGCTAAAATAAGTAAACAGGTAACGTCTATTGAATTAGACAGTCATCTATTCAACTTATCGTCAGAAAAA
TTAAAACTGAACATTCGTGTCACTTTAATTCACCAAGATATTCTACAGTTTCAATTCCCTAACAAACAGA
GGTATAAAATTGTTGGGAATATTCCTTACCATTTAAGCACACAAATTATTAAAAAAGTGGTTTTTGAAAG
CCATGCGTCTGACATCTATCTGATTGTTGAAGAAGGATTCTACAAGCGTACCTTGGATATTCACCGAACA
CTAGGGTTGCTCTTGCACACTCAAGTCTCGATTCAGCAATTGCTTAAGCTGCCAGCGGAATGCTTTCATC
CTAAACCAAAAGTAAACAGTGTCTTAATAAAACTTACCCGCCATACCACAGATGTTCCAGATAAATATTG
GAAGCTATATACGTACTTTGTTTCAAAATGGGTCAATCGAGAATATCGTCAACTGTTTACTAAAAATCAG
TTTCATCAAGCAATGAAACACGCCAAAGTAAACAATTTAAGTACCGTTACTTATGAGCAAGTATTGTCTA
TTTTTAATAGTTATCTATTATTTAACGGGAGGAAATAA
```

Blast Website



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

End of updates for BLAST+ version 4 databases (dbV4)

Start moving to the new version 5 databases!

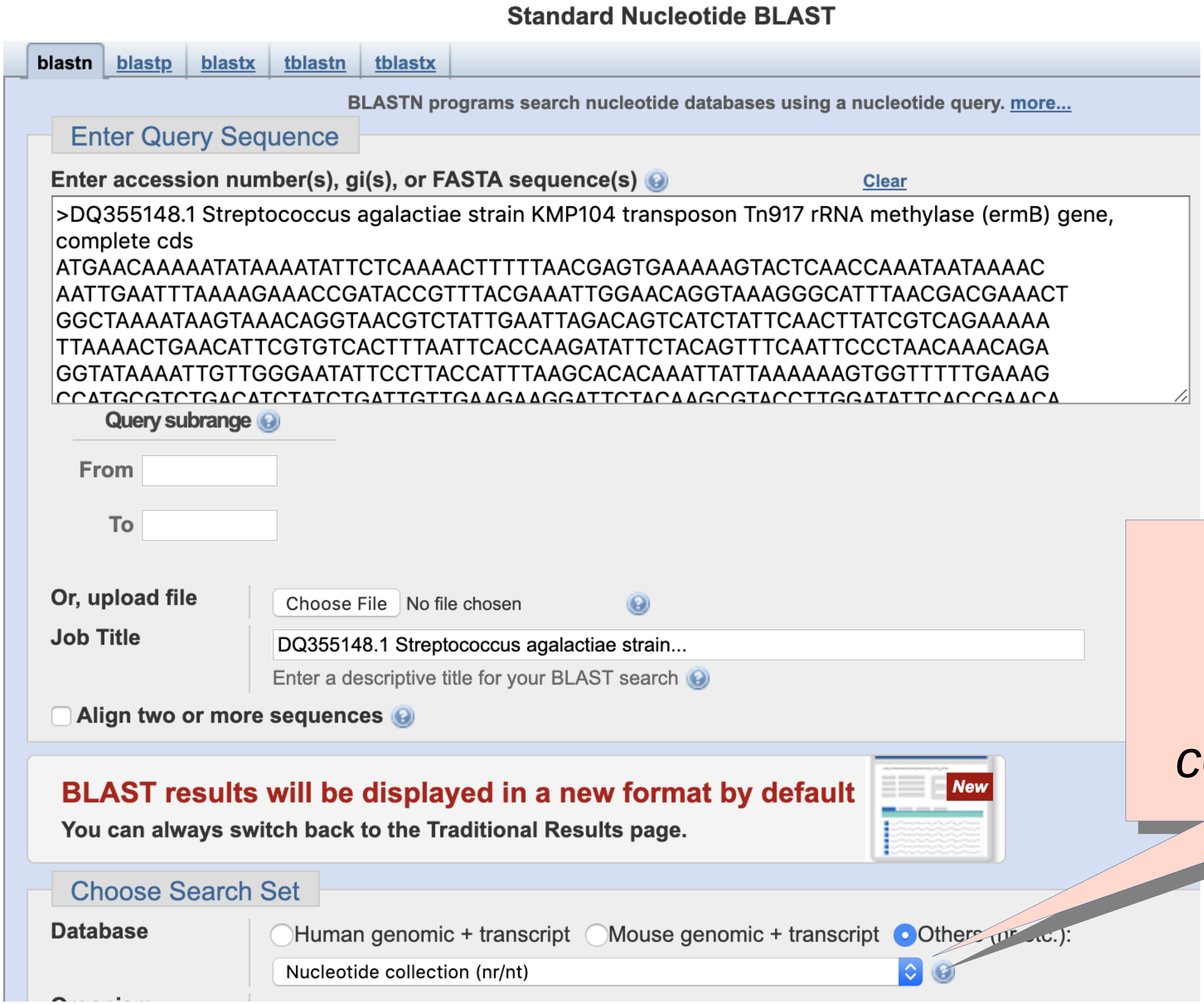
Fri, 27 Sep 2019 16:00:00 EST

More BLAST news...

Web BLAST

- https://blast.ncbi.nlm.nih.gov/Blast.cgi

Run The Query



Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
>DQ355148.1 Streptococcus agalactiae strain KMP104 transposon Tn917 rRNA methylase (ermB) gene,
complete cds
ATGAACAAAAATATAAAATATTCTCAAAACTTTTTAACGAGTGAAAAAGTACTCAACCAAATAATAAAAC
AATTGAATTTAAAAGAAACCGATACCGTTTACGAAATTGGAACAGGTAAAGGGCATTTAACGACGAAACT
GGCTAAAATAAGTAAACAGGTAACGTCTATTGAATTAGACAGTCATCTATTCAACTTATCGTCAGAAAAA
TTAAAACTGAACATTCGTGTCACTTTAATTCACCAAGATATTCTACAGTTTCAATTCCCTAACAAACAGA
GGTATAAAATTGTTGGGAATATTCCTTACCATTTAAGCACACAAATTATTAAAAAAGTGGTTTTTGAAAG
CCATGCGTCTGACATCTATCTGATTGTTGAAGAAGGATTCTACAAGCGTACCTTGGATATTCACCGAACA
```

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title: DQ355148.1 Streptococcus agalactiae strain...

Enter a descriptive title for your BLAST search

□ Align two or more sequences

BLAST results will be displayed in a new format by default

You can always switch back to the Traditional Results page.

Choose Search Set

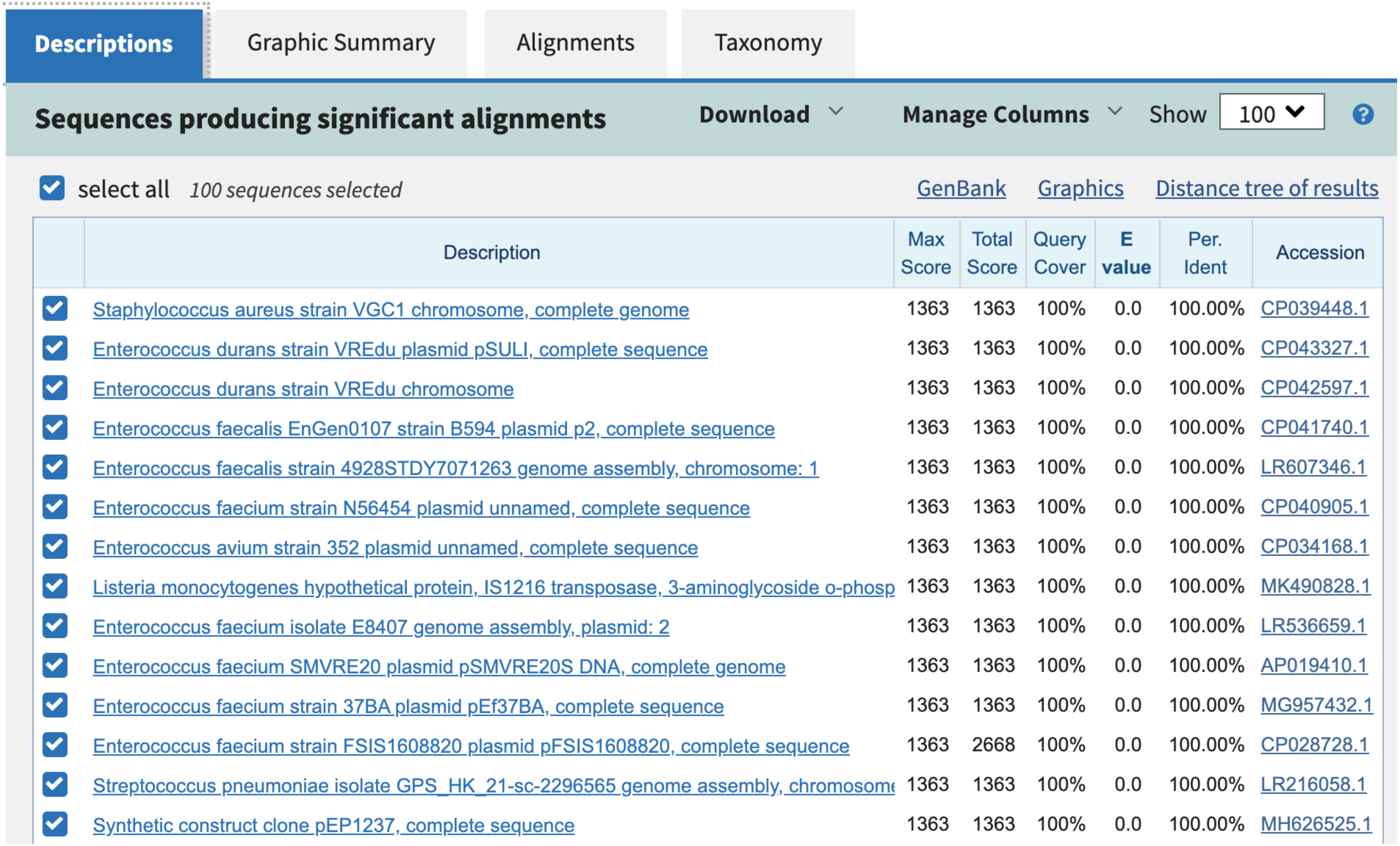
Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):

Nucleotide collection (nr/nt)

Organism

Use database: *Nucleotide collection (nr/nt)*

Results



Sequences producing significant alignments

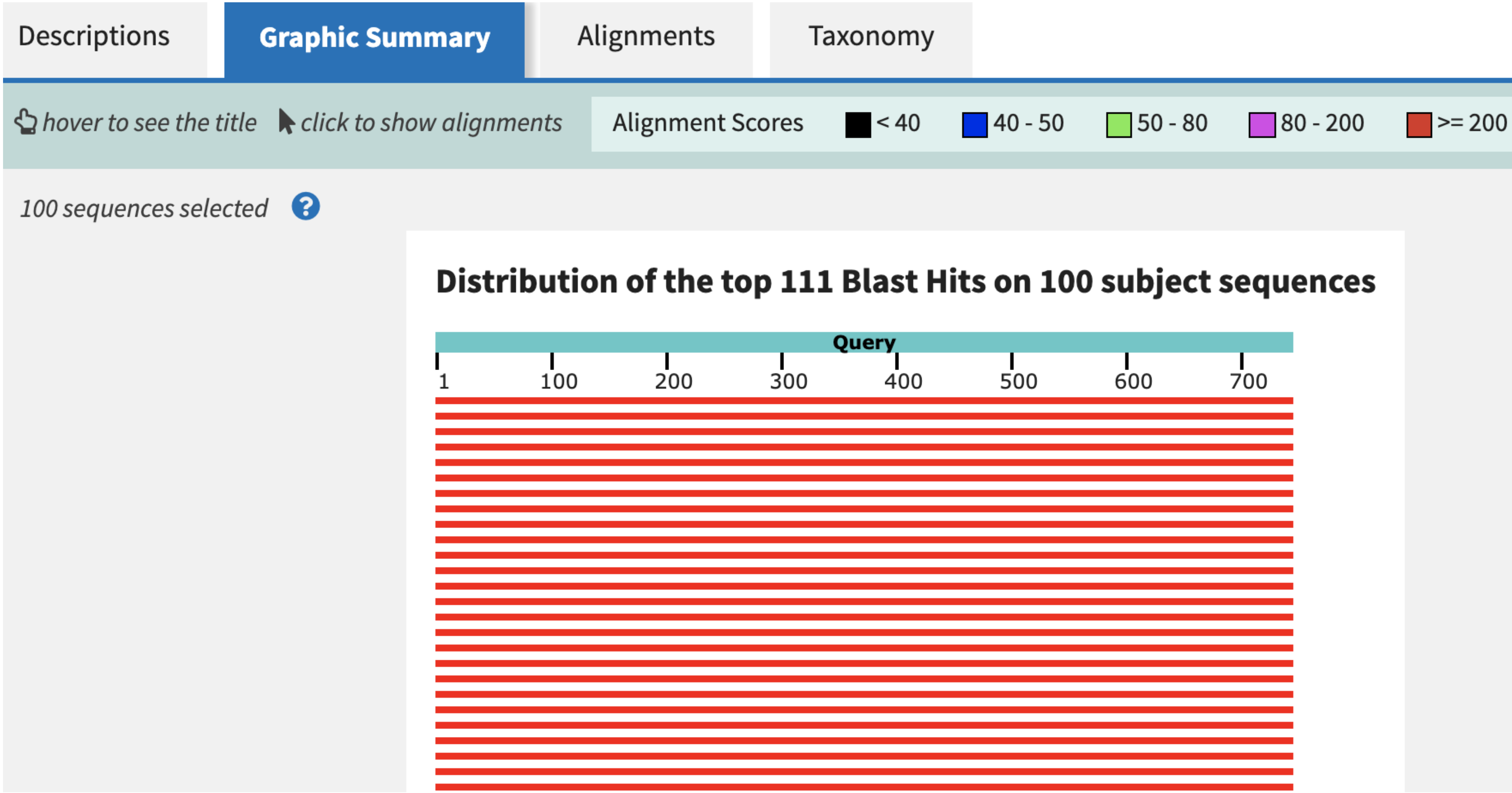
☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Staphylococcus aureus strain VGC1 chromosome, complete genome	1363	1363	100%	0.0	100.00%	CP039448.1
☑	Enterococcus durans strain VREdu plasmid pSULI, complete sequence	1363	1363	100%	0.0	100.00%	CP043327.1
☑	Enterococcus durans strain VREdu chromosome	1363	1363	100%	0.0	100.00%	CP042597.1
☑	Enterococcus faecalis EnGen0107 strain B594 plasmid p2, complete sequence	1363	1363	100%	0.0	100.00%	CP041740.1
☑	Enterococcus faecalis strain 4928STDY7071263 genome assembly, chromosome: 1	1363	1363	100%	0.0	100.00%	LR607346.1
☑	Enterococcus faecium strain N56454 plasmid unnamed, complete sequence	1363	1363	100%	0.0	100.00%	CP040905.1
☑	Enterococcus avium strain 352 plasmid unnamed, complete sequence	1363	1363	100%	0.0	100.00%	CP034168.1
☑	Listeria monocytogenes hypothetical protein, IS1216 transposase, 3-aminoglycoside o-phosp	1363	1363	100%	0.0	100.00%	MK490828.1
☑	Enterococcus faecium isolate E8407 genome assembly, plasmid: 2	1363	1363	100%	0.0	100.00%	LR536659.1
☑	Enterococcus faecium SMVRE20 plasmid pSMVRE20S DNA, complete genome	1363	1363	100%	0.0	100.00%	AP019410.1
☑	Enterococcus faecium strain 37BA plasmid pEf37BA, complete sequence	1363	1363	100%	0.0	100.00%	MG957432.1
☑	Enterococcus faecium strain FSIS1608820 plasmid pFSIS1608820, complete sequence	1363	2668	100%	0.0	100.00%	CP028728.1
☑	Streptococcus pneumoniae isolate GPS_HK_21-sc-2296565 genome assembly, chromosome	1363	1363	100%	0.0	100.00%	LR216058.1
☑	Synthetic construct clone pEP1237, complete sequence	1363	1363	100%	0.0	100.00%	MH626525.1

Scores

- **Max Score**
  - The score of the best matching segment for local alignment, not global
- **Total Score**
  - The total scores of all matching segments found (same as max score if there is only one matching segment)
- **Query Coverage**
  - The percentage of the query sequence that aligned to some part of the match.
- **E-Value**
  - A statistical measure evaluating how likely it is that a match this good could occur by chance. Lower e-scores indicate that both sequences are truly similar and are not similar by chance alone. Identical sequences have e-scores of zero.
- **Max Indent**
  - The percentage of nucleotides that are identical between the query and the target sequences within the matching regions.

Results

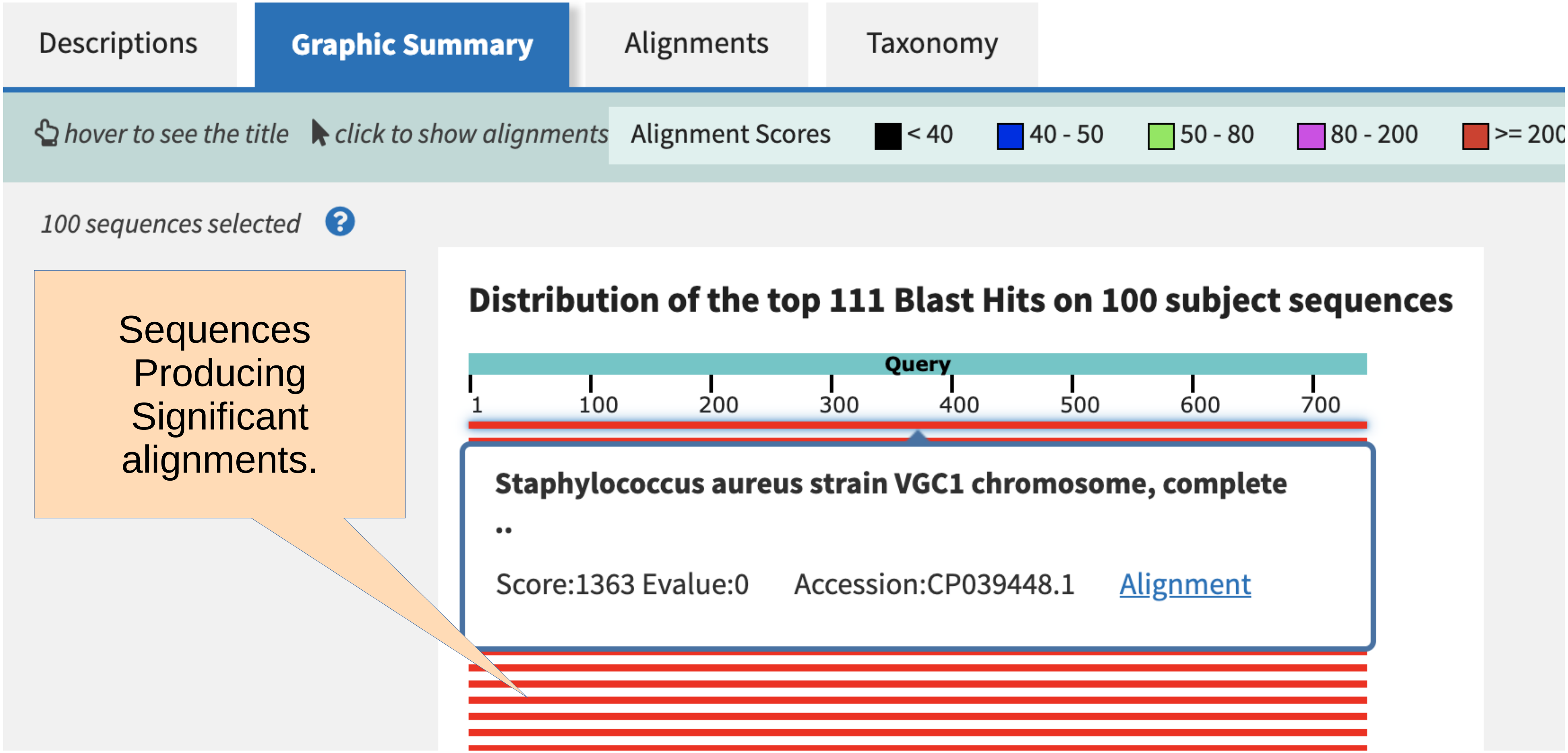


hover to see the title *click to show alignments*

100 sequences selected

Distribution of the top 111 Blast Hits on 100 subject sequences

Results



100 sequences selected ?

Sequences Producing Significant alignments.

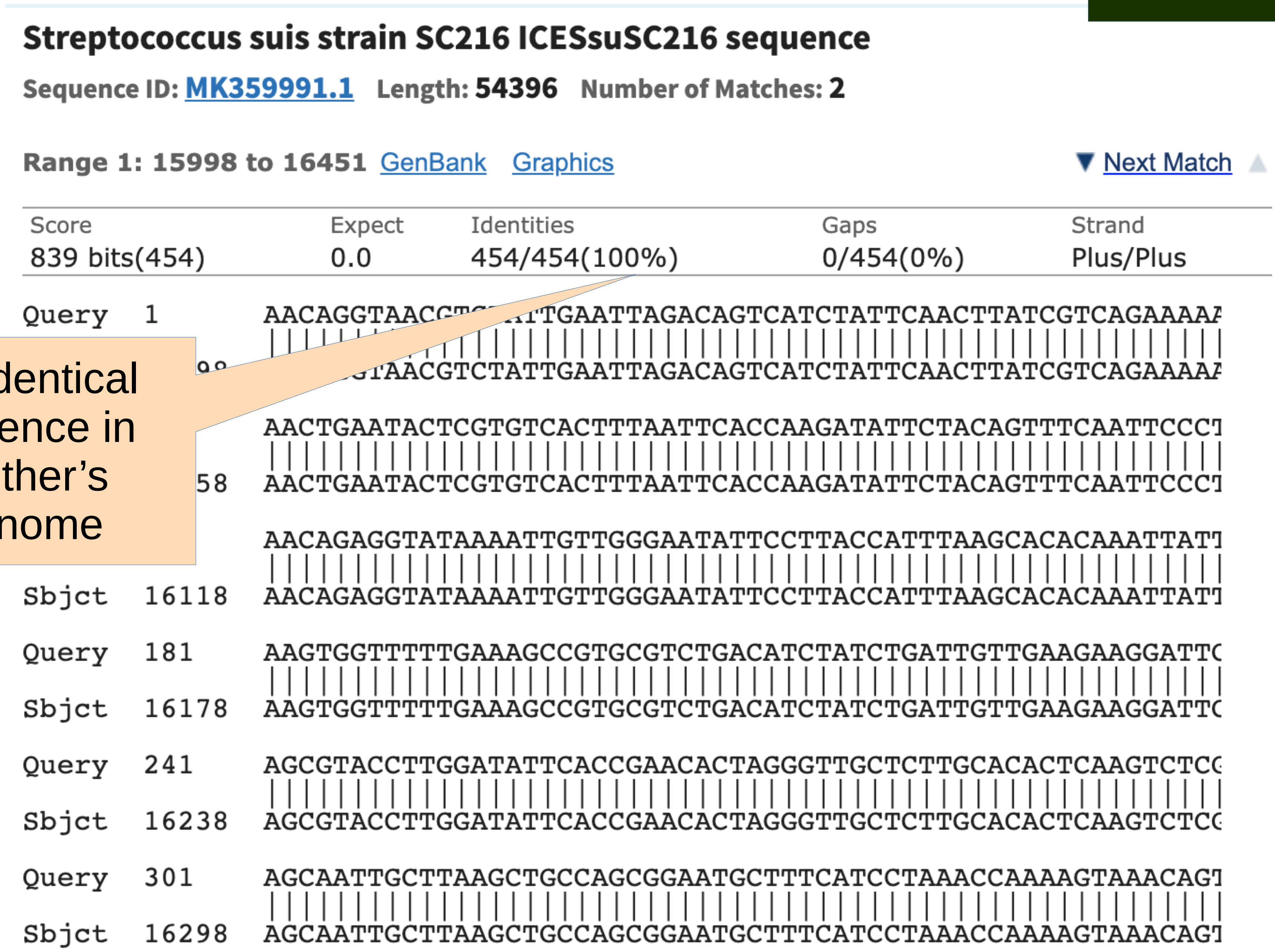
Distribution of the top 111 Blast Hits on 100 subject sequences

Staphylococcus aureus strain VGC1 chromosome, complete

..

Score:1363 Evalue:0 Accession:CP039448.1 Alignment

Results



Streptococcus suis strain SC216 ICESsuSC216 sequence

Sequence ID: MK359991.1 Length: 54396 Number of Matches: 2

Range 1: 15998 to 16451 GenBank Graphics ▼ Next Match ▲

Score	Expect	Identities	Gaps	Strand
839 bits(454)	0.0	454/454(100%)	0/454(0%)	Plus/Plus

```
Query  1      AACAGGTAACGTC  TTGAATTAGACAGTCATCTATTCAACTTATCGTCAGAAAA
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
         98     GTAACGTCTATTGAATTAGACAGTCATCTATTCAACTTATCGTCAGAAAA

              AACTGAATACTCGTGTCACTTTAATTCACCAAGATATTCTACAGTTTCAATTCCCT
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
         58   AACTGAATACTCGTGTCACTTTAATTCACCAAGATATTCTACAGTTTCAATTCCCT

              AACAGAGGTATAAAATTGTTGGGAATATTCCTTACCATTTAAGCACACAAATTATT
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16118  AACAGAGGTATAAAATTGTTGGGAATATTCCTTACCATTTAAGCACACAAATTATT

Query  181    AAGTGGTTTTTGAAAGCCGTGCGTCTGACATCTATCTGATTGTTGAAGAAGGATTC
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16178  AAGTGGTTTTTGAAAGCCGTGCGTCTGACATCTATCTGATTGTTGAAGAAGGATTC

Query  241    AGCGTACCTTGGATATTCACCGAACACTAGGGTTGCTCTTGCACACTCAAGTCTCG
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16238  AGCGTACCTTGGATATTCACCGAACACTAGGGTTGCTCTTGCACACTCAAGTCTCG

Query  301    AGCAATTGCTTAAGCTGCCAGCGGAATGCTTTCATCCTAAACCAAAAGTAAACAGT
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  16298  AGCAATTGCTTAAGCTGCCAGCGGAATGCTTTCATCCTAAACCAAAAGTAAACAGT
```

An Identical sequence in another's genome

Conclusions on HGT?

- Typically, researchers allow for a 95% similarity between genes found between *unrelated* organisms.
- Here, **we may conclude that HGT is a good hypothesis** but ***more research must be done*** to determine whether there was a chance for two organisms to be (physically) close enough to each other to share genetic material.

Did the bacteria Types meet somewhere? Like in a pond?

Blast is Cool!

Your Turn to Investigate!!!

- Investigate a gene of resistance: *ermA* (Accession number: **LT549456**)
- Questions:
  - What is the description of this gene? (hint: see Genbank record)
  - About how many other organisms appear to have traces of the same gene sequence?
  - What is the closest match? Which organism? What *e*-score? Conclusions?